DOCUMENT RESUME

ED 390 92[illegible] TM 024 427

AUTHOR Blais, Jean-Guy; Laurier, Michel
TITLE Methodological Considerations in Using DIMTEST To Assess Unidimensionality.
PUB DATE Apr 95
NOTE 16p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Francisco, CA, April 18-22, 1995).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Computer Software; *Evaluation Methods; Foreign Countries; French; *Item Response Theory; Second Language Learning; *Statistical Analysis; Student Placement; *Test Items
IDENTIFIERS Dimensionality (Tests); *DIMTEST (Computer Program); *Unidimensionality (Tests)

ABSTRACT

This study deals with the assessment of the unidimensionality of a set of items through the procedure developed by W. Stout and others (1991) and implemented in the computer program DIMTEST. This study examines a special feature of DIMTEST: the possibility for the user to assess the unidimensionality of a set of items by specifying a subset believed to form a homogeneous group of items. Data came from 3 subtests of an experimental version of a placement test in French as a second language, with samples of 698, 654, and 694 examinees. Results demonstrated the presence of essentially one dimension for the first subtest and of more than one dimension for the third subtest. These results also illustrate the applicability of the DIMTEST approach for studying dimensionality. The study further stresses the importance of considering unidimensionality as a multifaceted concept. An appendix describes the DIMTEST procedure in detail. (Contains 3 tables and 36 references.) (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 390 928

Methodological considerations in using DIMTEST to assess unidimensionality

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
This document has been reproduced as received from the person or organization originating it.
Minor changes have been made to improve reproduction quality.
Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
JEAN-GUY BLAIS
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

Jean-Guy Blais[1]
Michel Laurier

University of Montreal

April 1995

Presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

TM024427

[1] The research summarized here is collaborative in every respect and the listing of authors is alphabetical.

BEST COPY AVAILABLE

Methodological considerations in using DIMTEST to assess unidimensionality

The most popular mathematical models employed in Item Response Theory (IRT) all share the assumption that only one ability (or trait) is measured by the items under consideration. This assumption which is called *undimensionality* is not new in test construction. Although it has not always been formalized, unidimensionality has been a constant concern for those who analyse or interpret test results. This very restrictive assumption is never strictly met in the real world. Several factors related to the examinee, the test, the environment and interactions among these three factors always affect, to some extent, test performance. It is much more convenient and realistic to consider that the assumption of unidimensionality is met if it can be shown that a "dominant" dimension explains examinees' responses. However, even if it is worded in terms of a dominant dimension, the definition of unidimensionality is still fairly abstract and not very operational (Hambleton and Rovinelli, 1986).

Reckase (1990) proposed a distinction between psychological and statistical dimensionality and Henning (1992) emphasized the importance of this distinction with complex constructs such as second language competence. Psychological dimensionality would be related to the educational definition and statistical dimensionality would be the operational definition. The operational definition has been traditionaly used as a scientific "proxy" for the educational definition and many suggestions have been made in the attempt to devise a statistical method that would provide an efficient operational definition. The traditional statistical approach to the assessment of dimensionality is through factor-analytic methods (Zwick, 1987). For example, McDonald (1981, pp.14-15) said "it is reasonable to assert that a set of *n* tests or a set of *n* binary items is unidimensional if and only if it fits a nonlinear factor model with one common factor". In

this case, observing that a nonlinear factor model with one common factor fits the data would then be necessary to assess the presence of only one "dimension" and hence to conclude that the set of items is indeed unidimensional. With factor-analytic methods the psychological dimensionality is equated to the statistical dimensionality and one factor is synonymous with one dimension. However, McDonald (1981) has argued for the existence of minor components that are common to relatively few items at most. Then, it might be much more realistic, as we said before and as Humphreys (1984) stated, to assume that unidimensionality can only be approximated and that we better be looking for a "dominant" dimension than looking for only one dimension.

The importance of being able to show that a dominant dimension is responsible for examinees' performance is highlighted by a number of different simulation and real data studies that have shown that different IRT models provide good estimations of ability when such a single dominant trait exists. Estimation stability problems occur as other traits become more than minor. (See for these results: Reckase, 1979; Drasgow and Parsons, 1983; Doody-Bogan and Yen, 1983; Harrison, 1986; Wang, 1985, 1988; Blais, 1987; Greaud, 1988; Kim and Stout, 1993).

Hattie (1984, 1985) provided a quite comprehensive review of some of the statistical methods proposed to assess unidimensionality. The different methods could be classified according to their link with different test "theories" like the classical test theory (reliability, internal consistency and homogeneity index) or the item response theory (goodness of fit statistics), or with data reduction techniques like factor analysis (linear or nonlinear) and multidimensional scaling, or with structural equation modelling. Since Hattie's work, there has been other statistical proposals for an operational definition of unidimensionality. We can mention the work of Holland (1981), Rozenbaum (1984) and Holland and Rozenbaum (1986), and more recently, the procedure developed by Stout (1987, 1990) which was later refined by Nandakumar and Stout (1993).

This study deals with the assessment of the unidimensionality of a set of items with the Stout's procedure. The statistical procedure is the one implemented in the computer program DIMTEST (Stout, Nandakumar, Junker, Chang and Steidinger, 1991). The procedure is based on the concepts of essential dimensionality and essential independance (Stout, 1987, 1990). DIMTEST analyses a data set from of a group of N examinees who take an L item test. Each examinee produces a response vector that can be scored as 1s and 0s (dichotomously scored items). It is hypothesized that essential independance, with respect to a dominant trait, holds and that the item response functions are monotonic in regard of the same trait. The hypothesis is stated as follows: H_0: d_e =1 versus H_1: d_e >1, where d_e denotes the essential dimensionality of the space underlying a set of items (Nandakumar, 1993). The procedure is detailed in an appendix at the end of this paper. It implements the intuitive idea of splitting the set of items in different subtests and to compare them according to certain statistical properties they should have if "essential" unidimensionality was to hold.

The present study examines a special feature of DIMTEST: the possibility for the user to assess the unidimensionality of a set of items by specifying a subset of items believed to form a homogeneous group of items. More precisely, using this special feature of DIMTEST, we conducted a methodological inquiry of the unidimensionality of three subtests using some resampling techniques with real data sets.

This special feature included in DIMTEST is a very interesting one in that it allows the user, who could be a content expert very knowledgeable of the theoretical construct underlying items, to input different subsets of items in the assessment of unidimensionality. The user can compare the values of a single statistic about the contribution of other dimensions, computed either from a selection based on his expert judgment or from a selection based on a technical criteria (in this case, linear factor analysis). The possibility to use expert input is in accordance with recent approaches of test development that take into account the contribution of cognitive psychology.

(see in this direction Shepard, 1991; Bennett and Ward, 1993; Frederiksen, Mislevy and Bejar, 1993; Mislevy, 1993; Nichols 1994). With the development of a new generation of tests, the construct must include a sound theory of learning and a selection of content that take into account the underlying mental processes. In this context, it will become more appropriate for the user as an expert to be the one who decides which items are similar and which are not.

However, this special feature can have some drawbacks with real data sets (nothing is perfect in the real world of course). For some robustness considerations (Nandakumar and Stout, 1993), the optimal number of items to be included in a subset should be about 1/4 of the total number of items in the test and there should always be more than 4 items in the subset. For a 50 item test, this means that the subset should include about 12 items. The difficulty for the user shows up when he has to select these 12 items among the 50 items. In a real life, testing situation, it is more likely that the user will be able to divide the total number of items in, say, two or three subgroups of items believed to form homogeneous groups. For example, the user may be able to divide the 50 items of a test in three subgroups of 25, 18, and 7 items. Which subgroup should be used with DIMTEST? Is there a difference in the unidimensionality assessment if the first, second or third subgroup is chosen? Then, for subgroups containing more than 12 items, which items should be included in the subset? Finally, should the user decide to let DIMTEST choose the items, would the results be different than if he had selected the items himself? Our study address these questions that are of concern for the practician who will use DIMTEST in unidimensionality studies.

The data comes from an experimental version of a placement test in French as a Second Language containing 150 multiple-choice items with four options. Based on the nature of the tasks, the test was divided into three subtests of 50 items each. On the first subtest, the student reads a paragraph of approximately thirty words and is asked a multiple-choice question. We suspect that this test basically rests on a single trait that could be labelled "reading". On the second

subtest, the student reads the description, in English, of a current situation and must select the most appropriate statement among four French grammaticaly correct statements. The multiple aspects of appropriateness judgments suggest that unidimensionality may not hold in this subtest. The third subtest is a conventional "Fill-the-gap" exercise which focuses on grammar use and vocabulary. These two components may represent two dominant dimensions.

The test was administered to English-speaking Canadians enrolling in French summer classes in different colleges or universities participating in a national programme. The level of proficiency in French ranged from *absolute beginner* to *very advanced*. Due to the programme requirements, the examinees were fairly homogeneous in terms of linguistic and cultural background and in terms of age, education and socio-economic status. However some examinees who presented aberrant answer patterns were discarded to ensure the homogeneity of the sample. These examinees were detected using a reproducibility index obtained by creating a Guttman implicational scale of items and subjects. The first data set consisted of answers from 348 students who had completed the whole test. The data was then augmented with the answers of students who had written only one or two subtests. After deleting some examinees, we could then create a sample of 698 examinees for the first subtest, 654 for the second subtest and 694 for the third subtest. The three data sets thus formed are used in this study on the assessment of dimensionality with DIMTEST.

In line with the questions stated above, a specialist in second language teaching has been asked to divide each of the three subtests in homogeneous subgroups of items. The first subtest contained subgroups of 18, 28 and 4 items. The second subtest was divided in subgroups of 21 and 29 items. The third subtest contained subgroups of 22 and 28 items. For each subtest, the analysis were done using each subgroup of items (except for the 4 item subgroup of the first subtest).

The study was divided in three parts. *First*, we constructed, for each subgroup, 6 tailored sets of items: the 12 "most" content-related items and the 12 "least" content-related items, the 12 most and least difficult items, the 12 most and least discriminating items. The items were ordered in regard of their difficulty and discrimination after fitting a three-parameter logistic model to the data using BILOG (Mislevy and Bock, 1985). *Second*, we randomly generated 500 sets of 12 items for each subgroup. *Third*, we randomly generated 500 sets of 12 items out of 50 for each subtest not taking into account the content grouping. DIMTEST analyses were executed on each of the data sets thus so created. A regular DIMTEST analysis was also done based on the default factor analysis item grouping on each subtest.

For the data sets of the first part we have the following results as presented in Table 1. In the first subtest, there is statistical evidence ($\alpha = 0.05$) of more than one dimension only with the 18 item subgroup when we use the 12 most discriminating items. This finding echoes observations made by Nandakumar and Stout (1993) who modified later versions of DIMTEST to circumvent this statistical bias. In the second subtest, there is statistical evidence of more than one dimension only with the 21 item subgroup when we use the 12 most difficult items. Again, this result can be considered as an artefact of the procedure related to the same statistical bias. In the third subtest, there is statistical evidence of more than one dimension with the 22 item subgroup when we use the 12 most content-related items and with the 28 item subgroup when we use the 12 most discriminating items. The first result on this last subtest is probably an indication of the presence of more than one dimension whereas the second result can be due to the statistical bias.

		Difficulty		Discrimination		Content	
		+	-	+	-	+	-
Subtest #1	Item group #1	NO	NO	**YES**	NO	NO	NO
	Item group #2	NO	NO	NO	NO	NO	NO
Subtest #2	Item group #1	**YES**	NO	NO	NO	NO	NO
	Item group #2	NO	NO	NO	NO	NO	NO
Subtest #3	Item group #1	NO	NO	NO	NO	**YES**	NO
	Item group #2	NO	NO	**YES**	NO	NO	NO

Table 1: *Rejection of H_0 for the first part of the study*

Results for the data sets of the second part, are summarized in Table 2. For the first subtest, there is statistical evidence of more than one dimension in 4% of data sets (n=500) coming from the subgroup with 18 items and in 3.4% of data sets coming from the 28 item subgroup. For the second subtest, there is statistical evidence of more than one dimension in 15.4% of data sets coming from the subgroup with 21 items and in 2.8% of data sets coming from the 29 item subgroup. For the third subtest, there is statistical evidence of more than one dimension in 8.4% of data sets coming from the 22 item subgroup and in 17.8% of data sets coming from the 28 item subgroup.

	Item group #1	Item group #2
Subtest #1: Paragraph reading	4.0%	3.4%
Subtest #2: Appropriate statement	15.4%	2.8%
Subtest#3: "Fill-the-gap"	8.4%	17.8%

Table 2: *Rejection of H_0 for the second part of the study*

Finally, as shown in Table 3, the results for the third part data sets are as following. The first subtest gives statistical evidence of more than one dimension in 4.2% of the data sets. The second subtest in 5.4% of the data sets and the third subtest in 9.2% of the data sets. Regular DIMTEST analysis based on factor analysis item grouping does not raise statistical evidence of more than one dimension for any of the subtests.

	Random selections	FA selection
Subtest #1: Paragraph reading	4.2%	NO
Subtest #2: Appropriate satement	5.4%	NO
Subtest#3: "Fill-the-gap"	9.2%	NO

Table 3: *Rejection of H_0 for the third part of the study*

What could be concluded from the above results? We think that we can conclude for the presence of essentialy one dimension for the first subtest and for more than one dimension for the third subtest - even if the factor analysis based procedure didn't do so. It seems to us that the second subtest is more problematic. If the user had chosen the 12 items from the 29 item

subgroup, he would have had much more "chance" of getting statistical evidence for one dimension than if he had chosen items coming from the 21 item subgroup (2.8% against 15.4%).

It is possible for us to draw these conclusions because we have designed the tests and we, as content experts, know the process leading to items generation. The analysis is confirming a theoretically grounded item generation process. We already thought before doing any dimensionality analysis that there could be problems with the third subtest because vocabulary and grammar are two distinct constructs. We were also aware of some problems in relation with the multiple references of appropriateness judgments in the second subtest. On the other hand, mainly because of the importance of vocabulary knowledge in the first part, we expected a single dimension to emerge. Therefore, the DIMTEST results are confirming many of our expectations. It is interesting to note that the conclusions go in the same direction than those reached in a previous study done by the authors in which four different techniques were used to assess dimensionality of the three subtests (Blais and Laurier, in press). The four techniques were: structural equation modelling with LISREL (Jöreskog and Sörbom, 1983); full information factor analysis (Bock, Gibbons and Muraki, 1985) with TESTFACT ; single unidimensionality analysis with DIMTEST (Stout et al., 1991); a technique proposed by Bejar (1980) and based on item response theory calibration.

The present study should not be considered as final. At least three projects are on our research agenda. *First*, we should study, from a technical perspective, the multidimensionality of the second and third subtests with an index like Junker and Stout's ε (Junker and Stout, 1991). This additional study should give us some indication as to the appropriateness of using a unidimensional IRT model, if we want to do so, with multidimensional data. *Second*, whatever the results of the dimensionality study are, we should look at all the different sets of items that gave a large positive value of the T statistic. We would identify among these items those that regularly pop up. Then, we would examine these items very carefully to figure out under wich conditions a

large T value is produced and make a decision as to whether we should leave these items in the subtest or not. *Third*, we should look at all the generated set of items that gave a large negative value of the T statistic. The statistic T is said to be an index of unidimensionality only when it has a large positive value. Since it is computed from the difference between two quantities (T_L and T_B) that do not have the same meaning (see Stout, 1987 or the appendix in this paper) it would certainly be interesting to look at the items sets that gave a value of T_B much bigger than the value of T_L.

This study stresses again the importance of considering unidimensionality as a multi-faceted concept rather than from a yes or no perspective. It also emphasizes the role of the content expert as an important participant in the assessment of tests characteristics. We have to remind that the value of a formal test statistic is never in itself a sufficient justification for accepting a particular model (Goldstein, 1981). Finally, we think that useful methodological tools like resampling techniques should contribute more to the analysis of test data.

References

Bejar, I.I.(1980). A procedure for investigating the unidimensionality of achievement tests based on item parameters estimates. *Journal of Educational Measurement*, **17**, 283-296.

Bennett, R.E. and Ward, W.C. (Eds) (1993). ***Construction versus choice in cognitive measurement: Issues in constructed responses, performance testing, and portfolio assessment.*** Hillsdale, NJ: Lawrence Erlbaum.

Blais, J.-G. (1987). ***Effets de la violation du postulat d'unidimensionalité dans la théorie des réponses aux items.*** Unpublished doctoral thesis. Faculté des sciences de l'éducation. Université de Montréal. 181 pages.

Blais, J.-G., Laurier, M. (in press). The dimensionality of a language placement test from several analytical perspectives. ***Language Testing*, 12.**

Bock, R.D., Gibbons, R.D., and Muraki, E. (1985). ***Full information factor analysis.*** (MRC Report No 85-1). Chicago: National Opinion Research Center.

Doody-Bogan, E. and Yen, W.M. (1983). Detecting muitidimensionality and examinig its effect on vertical equating with the three paramater iogistic model. Paper presented at the annual meeting of the American Educational Research Association, Montreal, April 1983.

Drasgow, F. and Parsons, C.K. (1983). Application of unidimensional psychological item response theory models to multidimensional data. *Applied Psychological Measurement*, **7**, 189-99.

Frederiksen, N., Mislevy, R.J. and Bejar, I.I. (Eds) (1993). *Test theory for a new generation of tests*. Hillsdale: Erlbaum.

Goldstein, H. (1980). Dimensionality, bias, independance and measurement scale problems in latent trait test score models. *British Journal of Mathematical and Statistical Psychology*, **33**, 234-246.

Greaud, V.A. (1988). Some effects of applying unidimensional IRT to multidimensional tests. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, April.

Hambleton, R.K. and Rovinelli, R.J. (1986). Assessing the dimensionality of a set of test items. *Applied Psychological Measurement*, **10**, 287-302.

Harrison, D.A. (1986). Robustness of IRT parameter estimation to violations of the unidimensionality assumption. *Journal of Educational Statistics*, **11**, 91-115.

Hattie, J.A. (1984). An empirical study of various indices for determining unidimensionality. *Multivariate Behavioral Research*, **19**, 49-78.
Hattie, J.A. (1985). Methodology review: Assessing unidimensionality ot test and items. *Applied Psychological Measurement*, **9**, 139-64.

Henning, G. (1992). Dimensionality and construct validity of language tests. *Language Testing*, **9**, 1-11.

Holland, P.W. (1981). When are item response models consistent with observed data? *Psychometrika*, **46**, 79-92.

Holland, P.W. and Rozenbaum, P.R. (1986). Conditional association and unidimensionality in motone latent trait variable models. *Annals of Statistics*, **14**, 1523-1543.

Humphreys, L. (1984). *A theoretical and empirical study of the psychometric assessment of psychological test dimensionality and bias* (ONR Research Proposal). Washington, DC: Office of Naval Research.

Jöreskog, K.W. and Sörbom, D. (1983). *LISREL User's guide*. Department of Statistics, University of Uppsala.

Junker, B.W. and Stout, W.F. (1991). Robustness of ability estimation when multiple trait is present. Paper presented at the 1991 International Symposium on Modern Theories in Measurement: Problems and Issues, Montebello, Quebec.

Kim, H.R. and Stout, W.F. (1993). A robustness study of ability estimation in the presence of latent trait multidimensionality using the Junker/Stout index of dimensionality. Paper presented at the annual meeting of the American Educational Research Association, Atlanta, April.

McDonald, R.P. (1981). The dimensionality of tests and items. *British Journal of Mathematical and Statistical Psychology*, **27**, 82-89.

Mislevy, R.J. (1993). Test theory reconceived. Paper presented at the annual meeting of the American Educational Research Association, Atlanta, April.

Mislevy, R.J. and Bock, R.D. (1986). *PC-BILOG: Item analysis and test scoring with binary logistic models*. Mooresville, IN: Scientific Software.

Nandakumar, R. (1993). Assessing dimensionality of real data. *Applied Psychological Measurement*, **17**, 29-38.

Nandakumar, R. and Stout, W. (1993). Refinements of Stout's procedure for assessing latent trait unidimensionality. *Journal of Educational Statistics*, **18**, 41-68.

Nichols, P.D. (1994). A framework for developing cognitively diagnostic assessment. *Review of Educational Research*, **64**, 575-603.

Reckase, M.D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational Statistics*, **4**, 207-230.

Reckase, M.D. (1990). Unidimensional Data from Multidimensional Tests and Multidimensional Data from Unidimensional Tests. Paper presented at the annual meeting of the American Educational Research Association, Boston, April.

Rozenbaum, P.R. (1984). Testing the conditional independance and monotonicity assumptions of item response theory. *Psychometrika*, **49**, 425-436.

Shepard, L. A. (1991). Psychometricians' Beliefs About Learning. *Educational Researcher*, vol.20, no 7, 2-16.

Stout, W. (1987). A non-parametric approach for assessing latent trait unidimensionality. *Psychometrika*, **52**, 589-617.

Stout, W. (1990). A new item response theory modeling approach with applications to unidimensional assessment and ability estimation. *Psychometrika*, **55**, 293-326.

Stout,W., Nandakumar, R., Junker, B., Chang, H. and Steidinger, D. (1991). *DIMTEST and TESTSIM, Programs for Dimensionality Testing and Test Simulation*. University of Illinois at Urbana-Champaign, Department of Statistics.

Wang, M. (1985). *Fitting a unidimensional model to multidimensional item response data: The effects of latent space misspecification on the application of IRT*. Unpublished Manuscript, University of Iowa.

Wang, M. (1988). Measurement bias in the application of a unidimensional model to multidimensional item-response data. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, April.

Zwick, R. (1987). Assessment of the dimensionality of NAEP Year 15 reading data. In A. E. Beaton, *Implementing the new design: The NAEP 1983-1984 Technical Report*. Report no: 15-TR-20. National Assessment of Educational Progress.

APPENDIX

The procedure described by Stout (1987) and ameliorated by Nandakumar and Stout (1993) uses a statistic T to test the hypothesis that there is only one dimension, H_0 : d=1, against the alternative that there is more than one dimension H_1 : d>1.

Observations can be represented by $\{U_{ij}\}$ where i, $1 \leq i \leq n + M$, indexes items and j, $1 \leq j \leq J$, indexes examinees. In the present case observations are response vectors of 0's and 1's with 1 denoting a correct response to an item and 0 denoting an incorrect response. The differents steps to calculate T when the number of examinees is small (under 2000 are the following):

Step 1: Split test into partitioning and assessment subtests.

The N test items are split into a short assessment subtest of lenght M and a long partioning subtest of lenght n. For some robustness considerations it is preferable that $4 \leq M \leq N/4$ (Nandakumar and Stout, 1993).

The M items can be chosen along two strategies. First they can represent a homogeneous set of items in the opinion of an expert. Secondly, they load most heavily positively or negatively on the second extracted factor of a principal axis factor analysis (with no rotation) of the tetrachoric corrlation coefficients with maximum observed correlations in each column used in place of communalities.

Step 2: Assign examinees to subgroups.

The examinees are assigned to different subgroups according to their differing partitioning subtest scores. It is required that each subgroup has a "large" number of examinees. All subgroups with less than J_{min} examinees are deleted and a number of $J_{min} \geq 20$ is recommended to maintain close agreement with the asymptotic theory.

Step 3: Compute the "usual" variance estimate for the k-th subgroup.

Let U_{ijk} indicate the correctness of the response of the j-th examinee from subgroup k to the i-th assessment item.
Let $J_k = J_k^{(n)}$ denote the number of examinees of subgroup k and $K = K^{(n)} \leq n - 1$ denote the number of subgroups.

Let $$Y_j^{(k)} = \sum_{i=1}^{M_k} \frac{U_{ijk}}{M}$$

denote the assessment subtest score of the j-th examinee from subgroup k.

Let $$\bar{Y}^{(k)} = \sum_{j=1}^{J_k} \frac{Y_j^{(k)}}{J_k}$$

denote the average examinee assessment subtest score for subgroup k.

Finally, let $$\hat{\sigma}_k^2 = \sum_{j=1}^{J_k} \frac{\left(Y_j^{(k)} - \bar{Y}^{(k)}\right)^2}{J_k}$$

denote the variance estimate of examinee assessment subtest scores in subgroup k.

Step 4: Compute the "unidimensional" variance estimate for the k-th subgroup.

Let $\hat{\sigma}^2_{U,k} = \sum_{i=1}^{M} \frac{\hat{p}_i^{(k)} - \left(1 - \hat{p}_i^{(k)}\right)}{M^2}$,

where $\hat{p}_i^k = \sum_{j=1}^{J_k} \frac{U_{ijk}}{J_k}$,

denote the "unidimensional" variance estimate for subgroup k.

Step 5: Normalize and combine the different subgroup variance estimates to form the statistic.

Let $\hat{\mu}_{4,k} = \sum_{j=1}^{J_k} \frac{\left(Y_j^{(k)} - \bar{P}^{(k)}\right)^4}{J_k}$, $\hat{\delta}_{4,k} = \sum_{i=1}^{M} \hat{p}_i^{(k)}\left(1 - \hat{p}_i^{(k)}\right)\left(1 - 2\hat{p}_i^{(k)}\right)^2$

and $S_k'^2 = \frac{\left[\left(\hat{\mu}_{4,k} - \hat{\sigma}_k^4\right) + \delta_{4,k}/M^4\right]}{J_k}$.

Let the statistic T_L be: $T_L = \frac{1}{K^{1/2}} \sum_{k=1}^{K} \frac{X_k}{S_k'}$

where $X_k = \hat{\sigma}_k^2 - \hat{\sigma}_{U,k}^2$.

Each X_k measures non-unidimensionality in the sense that $X_k = 0$ when $d_E = 1$ and $X_k > 0$ on average when $d_E > 1$.

Step 6: Correct for statistical bias.

Select a set of M items from that would otherwise be the n partitioning subtest items such that the selected items have an item difficulty distribution as similar as possible to that of the assessment subtest. Compute T_B like T_L but in using the assessment subtest 2 items. The bias corrected statistic is defined by:

$$T = \frac{\left(T_L - T_B\right)}{\sqrt{2}} .$$

Since assessment subtests are selected to have a difficulty distribution similar to that of assessment subtest 1 this allows T_B to compensate for the influence of item difficulty bias on T_L.

Step 7: Perform the test for unidimensionality for J small; that is $J \leq 2000$.

Reject H_0: $d = 1$ if $T > Z_\alpha$ where Z_α is the upper $100(1-\alpha)$ percentile for a standard normal distribution, α being the desired level of significance.